Deciphering Multi-way Interactions in the Human Genome

Stephen Lindsly[1,†], Can Chen[2,3,†], Sam Dilworth[4], Sivakumar Jeyarajan[1], Walter Meixner[1], Cooper Stansbury[1], Anthony Cicalo[1], Nicholas Beckloff[5], Charles Ryan[6,7], Amit Surana[8], Max Wicha[9], Gilbert S. Omenn[1], Lindsey Muir[1], and Indika Rajapakse[1,2,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109 USA
[2]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 USA
[3]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor 48109 USA
[4]iReprogram, Ann Arbor, MI 48105 USA
[5]Oxford Nanopore Technologies, Oxford, OX4 4DQ UK
[6]Medical Scientist Training Program, University of Michigan, Ann Arbor, MI 48109 USA
[7]Program in Cellular and Molecular Biology, University of Michigan, Ann Arbor, MI 48109 USA
[8]Raytheon Technologies Research Center, East Hartford, CT 06108 4 warnings USA
[9]Department of Hematology/Oncology, University of Michigan, Ann Arbor, MI 48109 USA
[†]These authors contributed equally to this work.
[*]To whom correspondence should be addressed (indikar@umich.edu).

Abstract

Chromatin architecture, a key regulator of gene expression, is inferred through chromatin contacts. However, classical analyses of chromosome conformation data do not preserve multi-way relationships. Here we use long sequencing reads to map genome-wide multi-way contacts and investigate higher order chromatin organization of the human genome. We use the theory of hypergraphs for data representation and analysis, and quantify higher order structures in primary human fibroblasts and B lymphocytes. Through integration of multi-way contact data with chromatin accessibility, gene expression, and transcription factor binding data, we introduce a data-driven method to extract transcriptional clusters.

1 Introduction

Structural features of the genome are integral to regulation of gene expression and corresponding outcomes in cellular phenotype [1, 2, 3]. The biochemical technique of genome wide chromosome conformation capture (Hi-C) enables large scale investigation of genome architecture through capture of all adjacent loci. Through analysis of pairwise interactions, multi-way interactions may be inferred, but not captured directly [4]. Recently, an extension of Hi-C enabled preservation and sequencing of long reads (e.g. Pore-C) [5], which can be used to unambiguously identify multiple adjacent loci. These data will undoubtedly clarify higher order structures in the genome, but new frameworks are required for analysis and representation of the multidimensional data captured in this assay.

Here we collected Pore-C data from two cell types and constructed hypergraphs to represent the multidimensional relationships of multi-way contacts among loci. Hypergraphs are similar to graphs, but have hyperedges instead of edges, with each hyperedge connecting any number of nodes instead of two nodes [6, 7, 8]. In our hypergraph framework, nodes are genomic loci, and hyperedges are multi-way contacts among loci. We used incidence matrices to represent hypergraphs, where matrix rows represent genomic loci and columns are individual hyperedges. This representation enables quantitative measurements of genome organization through hypergraph entropy, comparison of different cell types through hypergraph distance, and identification of functionally important multi-way contacts in multiple cell types. In addition, we integrate Pore-C data with multiple other data modalities to identify biologically relevant multi-way interactions.

We use the following definitions. **Entropy**: a measure of structural order in the genome. **Hyperedge**: an extension of edges where each edge can contain any number of nodes. **Hypergraph**: an extension of graphs containing multiple hyperedges. **Incidence matrix**: a representation for hypergraphs where rows are nodes and columns are hyperedges. **Transcription cluster**: a group of genomic loci that colocalize for efficient gene transcription.

2 Results

2.1 Capturing Multi-way Contacts

We conducted Pore-C experiments using adult human dermal fibroblasts and obtained additional publicly available Pore-C data from B lymphocytes [5].) The experimental protocol for Pore-C is similar to Hi-C, where DNA is cross-linked to histones, restriction digested, and adjacent ends are ligated together followed by sequencing (Figure 1A). Alignment of Pore-C long reads to the genome enables fragment identification and construction of multi-way contacts (Figure 1B).

We use hypergraphs to represent multi-way contacts, where individual hyperedges contain at least two loci (Figure 1C, left). Hypergraphs provide a simple and concise way to depict multi-way contacts and allow for abstract representations of genome structure. Computationally, we represent multi-way contacts as incidence matrices (Figure 1C, right). For Hi-C data, adjacency matrices are useful for assembly of pairwise genomic contacts. However, since rows and columns represent individual loci, adjacency matrices cannot be used for multi-way contacts in Pore-C data. In contrast, incidence matrices permit more than two loci per contact and provide a clear visualization of multi-way contacts. Multi-way contacts can also be decomposed into pairwise contacts, similar to Hi-C, by extracting all combinations of loci (Figure 1D).

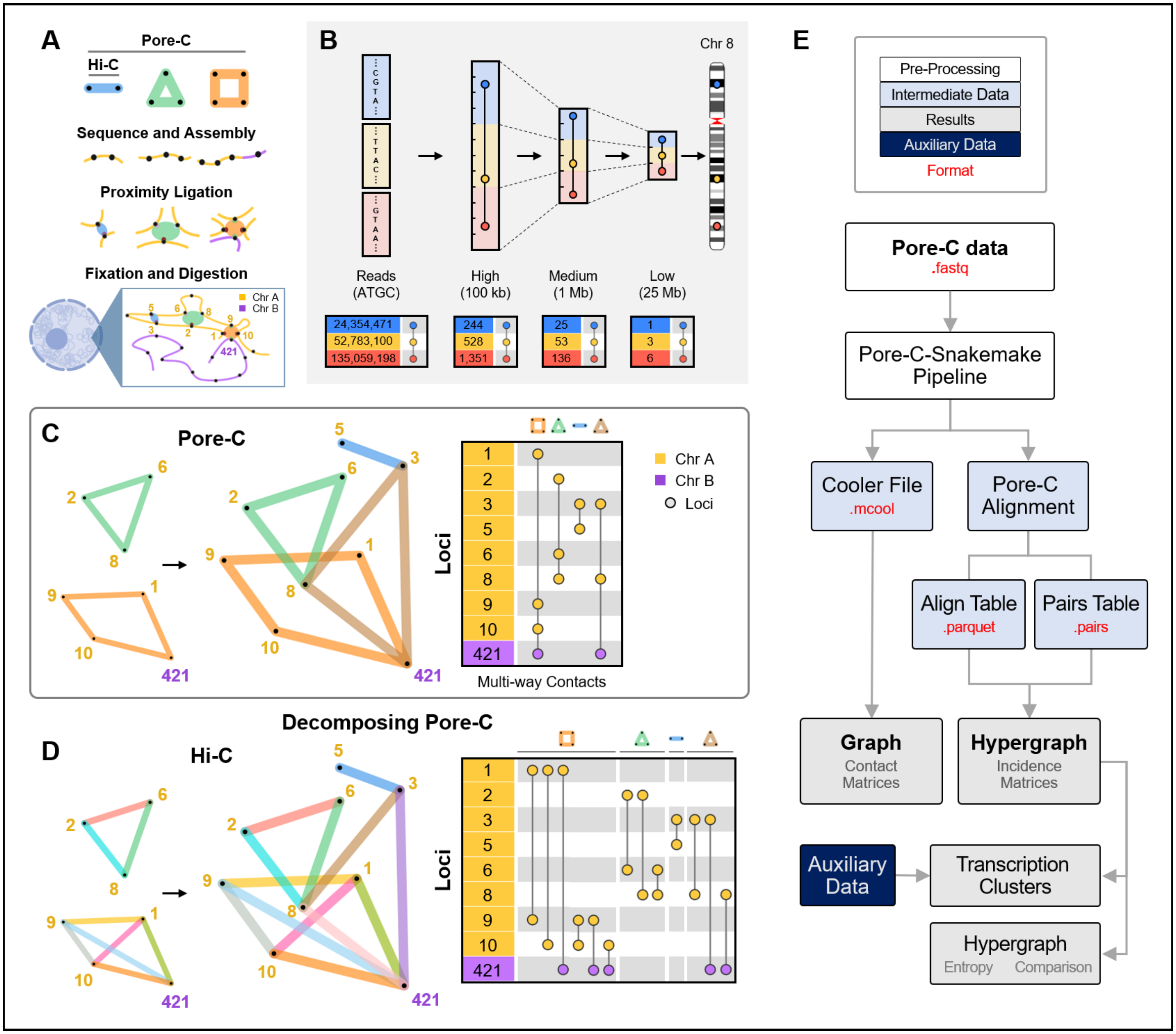

Figure 1: Pore-C experimental and data workflow. (A) The Pore-C experimental protocol, which captures pairwise and multiway contacts (Materials and Methods). (B) Representation of multi-way contacts at different resolutions (top). Incidence matrix visualizations of a representative example from Chromosome 8 in human fibroblasts at each resolution (bottom). The numbers in the left columns represent the location of each genomic locus present in a multi-way contact, where values are either the chromosome base-pair position (read-level) or the bin into which the locus was placed (binning at 100 kb, 1 Mb, or 25 Mb). (C) Hypergraph representation of Pore-C contacts (left) and an incidence matrix (right) of four sets of multi-way contacts within (yellow-to-yellow) and between (yellow-to-purple) chromosomes. Contacts correspond to examples from part A. The numbers in the left column represent a bin in which a locus resides. Each vertical line represents a multi-way contact, with nodes at participating genomic loci. (D) Multi-way contacts can be decomposed into pairwise contacts. Decomposed multi-way contacts can be represented using graphs (left) or incidence matrices (right). Contacts correspond to examples from parts A and C. (E) Flowchart overview of the computational framework. File type format descriptions are in Table S1.

2.2 Decomposing Multi-way Contacts

From our Pore-C experiments using adult human dermal fibroblasts and additional publicly available Pore-C data from B lymphocytes, we constructed hypergraphs as multiple resolutions (read level, 100 kb, 1 Mb, and 25 Mb) [5]. We first analyzed individual chromosomes at 100 kb resolution, and decomposed the multi-way contacts into their pairwise counterparts to identify topologically associated domains (TADs, *Materials and Methods*). Examples of TADs from Chromosome 22 for fibroblasts and B lymphocytes can be seen in Figures 2 and S1, respectively.

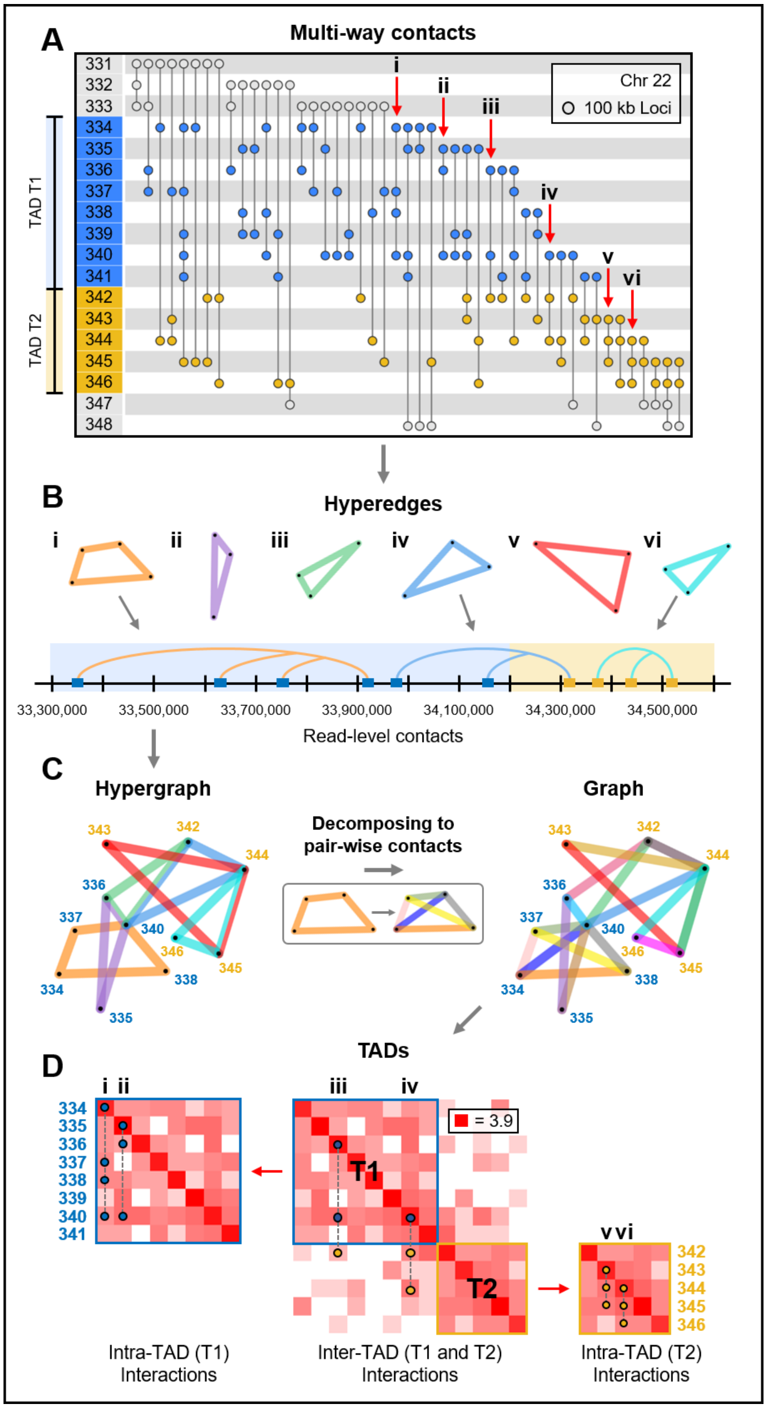

Figure 2: Local organization of the genome. (A) Incidence matrix visualization of a region in Chromosome 22 from fibroblasts (V1-V4). The numbers in the left column represent genomic loci in 100 kb resolution, vertical lines represent multi-way contacts, where nodes indicate the corresponding locus' participation in this contact. The blue and yellow regions represent two TADs, T1 and T2. The six contacts, denoted by the labels i-vi, are used as examples to show intra- and inter-TAD contacts in B, C, and D. (B) Hyperedge and read-level visualizations of the multi-way contacts i-vi from A. Blue and yellow rectangles (bottom) indicate which TAD each loci corresponds to. (C) A hypergraph is constructed using the hyperedges from B (multiway contacts i-vi from A). The hypergraph is decomposed into its pairwise contacts in order to be represented as a graph. (D) Contact frequency matrices were constructed by separating all multi-way contacts within this region of Chromosome 22 into their pairwise combinations. TADs were computed from the pairwise contacts using the methods from [9]. Example multi-way contacts i-vi are superimposed onto the contact frequency matrices. Multi-way contacts in this figure were determined in 100 kb resolution after noise reduction, originally derived from read-level multi-way contacts (*Materials and Methods*).

2.3 Chromosomes as Hypergraphs

To gain a better understanding of genome structure with multi-way contacts, we constructed hypergraphs for entire chromosomes in 1 Mb resolution. We show the incidence matrix of Chromosome 22 as an example in Figure 3A. In Figure 3B, we show Chromosomes 22's distribution of 1 Mb contacts at multiple orders (2-way contacts, 3-way contacts, etc). Figure 3C highlights the most common intra-chromosome contacts within Chromosome 22 using multi-way contact "motifs", which we use as a simplified way to show hyperedges. Figure 3D shows a zoom-in of a 3-way contact to highlight how a low resolution multi-way contact can contain many contacts at higher resolutions. Figure 3E visualizes the multi-way contacts contained in Figure 3D as a hypergraph.

We also identified multi-way contacts that contain loci from multiple chromosomes. These inter-chromosomal multi-way contacts can be seen in 1 Mb resolution in Figure 3F and in 25 Mb resolution for both fibroblasts and B lymphocytes in Figure 4A and 4B, respectively. Figure 4 gives a summary of the entire genome's multi-way contacts, by showing the most common intra- and inter-chromosomal multi-way contacts across all chromosomes. We highlight examples of multi-way contacts with loci that are contained within a single chromosome ("intra only"), spread across unique chromosomes ("inter only"), and a mix of both within and between chromosomes ("intra and inter"). Finally, we found the most common inter-chromosomal multi-way contacts across all chromosomes, which we summarize with five example chromosomes in Figure 5 using multi-way contact motifs.

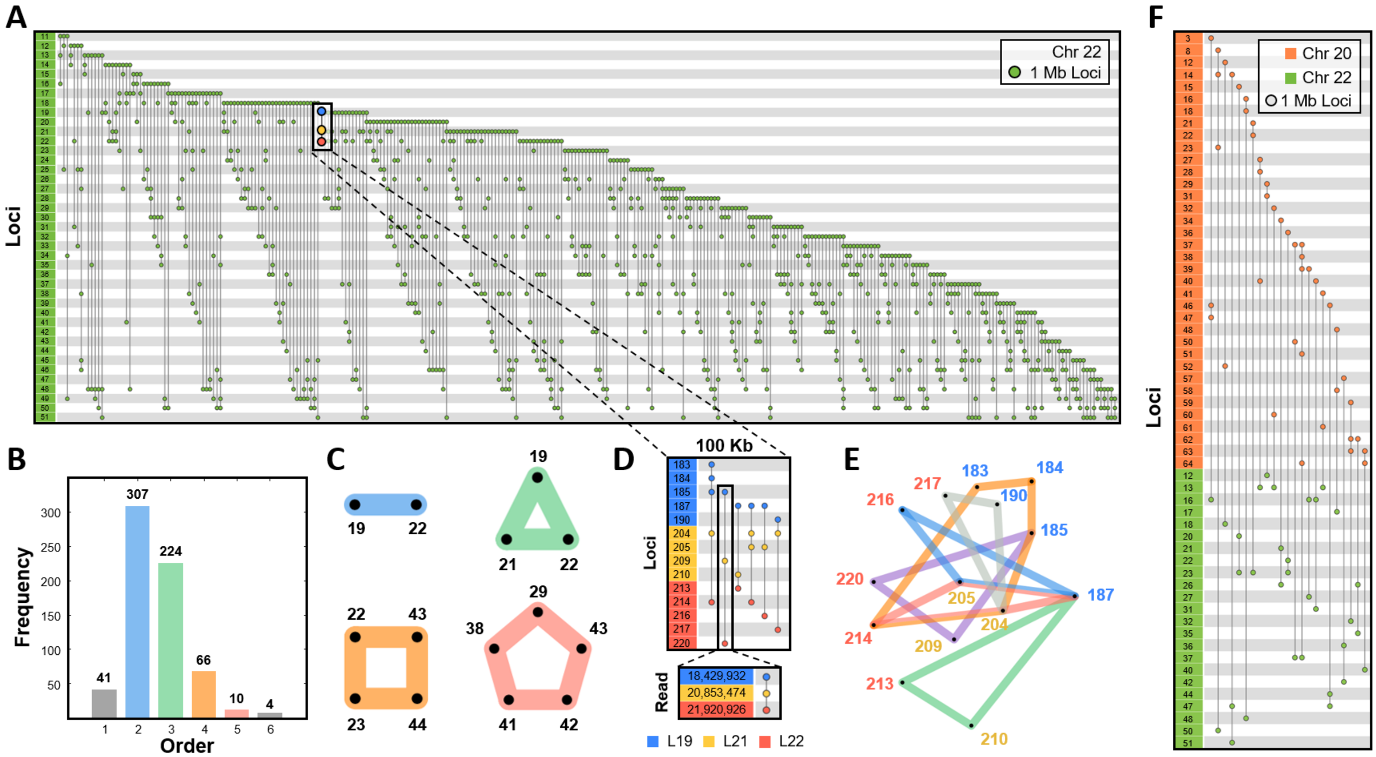

Figure 3: Patterning of intra- and inter-chromosomal contacts. (A) Incidence matrix visualization of Chromosome 22 in fibroblasts. The numbers in the left column represent genomic loci in 1 Mb resolution. Each vertical line represents a multi-way contact, in which the nodes indicate the corresponding locus' participation in this contact. (B) Frequencies of Pore-C contacts in Chromosome 22. Bars are colored according to the order of contact. Blue, green, orange, and red correspond to 2-way, 3-way, 4-way, and 5-way contacts. (C) The most common 2-way, 3-way, 4-way, and 5-way intra-chromosome contacts within Chromosome 22 are represented as motifs, color-coded similarly to B. (D) Zoomed in incidence matrix visualization in 100 kb resolution shows the multi-way contacts between three 1 Mb loci L19 (blue), L21 (yellow), and L22 (red). An example 100 kb resolution multi-way contact is zoomed to read-level resolution. (E) Hypergraph representation of the 100 kb multi-way contacts from D. Blue, yellow, and red labels correspond to loci L19, L21, and L22, respectively. (F) Incidence matrix visualization of the inter-chromosomal multi-way contacts between Chromosome 20 (orange) and Chromosome 22 (green) in 1 Mb resolution. Within this figure, all data are from one fibroblast sequencing run (V2) and multi-way contacts were determined after noise reduction at 1 Mb or 100 kb resolution accordingly (*Materials and Methods*).

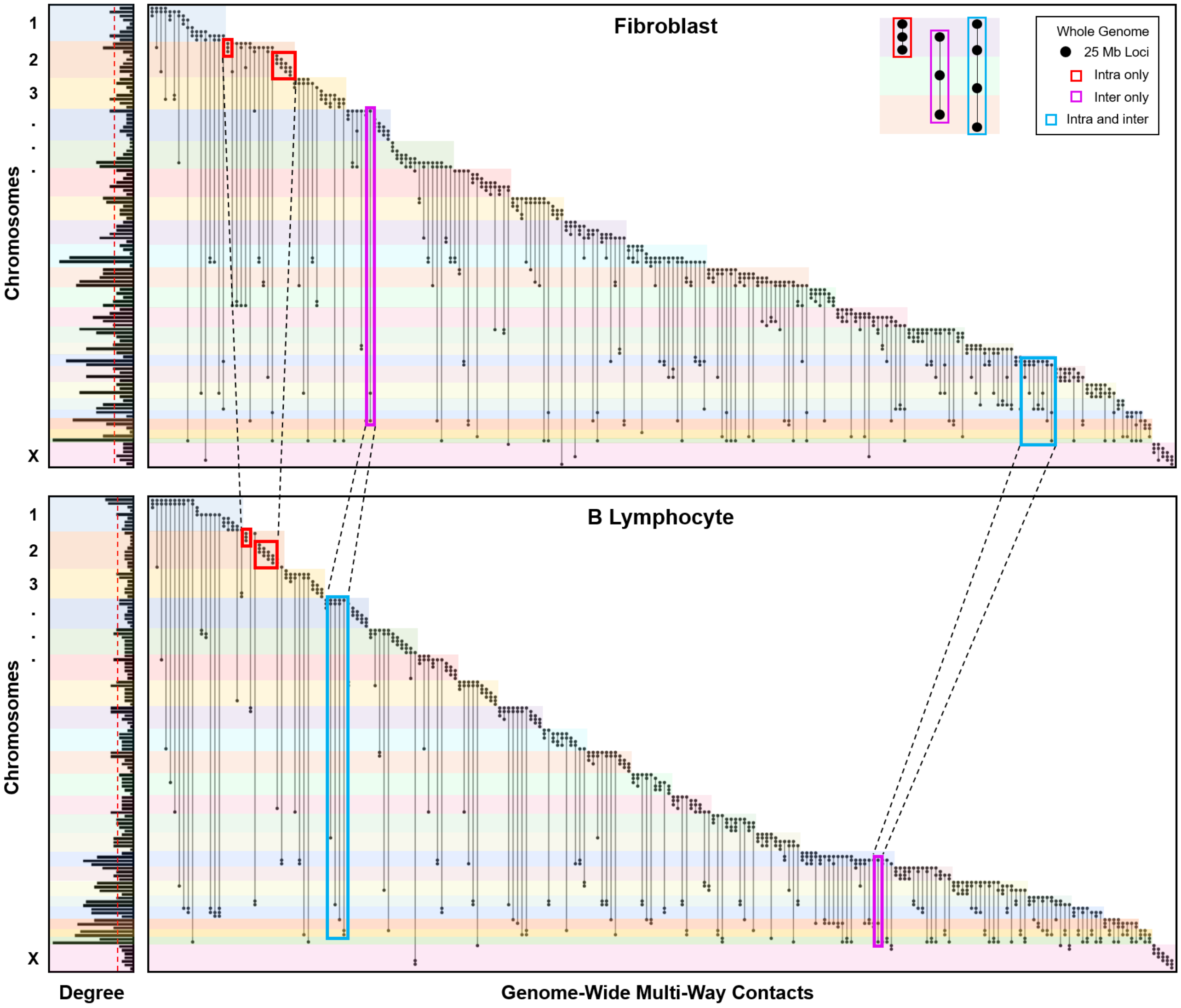

Figure 4: Genome-wide patterning of multi-way contacts. Incidence matrix visualization of the top 10 most common multi-way contacts per chromosome. Matrices are constructed at 25 Mb resolution for both fibroblasts (top, V1-V4) and B lymphocytes (bottom). Specifically, 5 intra-chromosomal and 5 inter-chromosomal multi-way contacts were identified for each chromosome with no repeated contacts. If 5 unique intra-chromosomal multi-way contacts are not possible in a chromosome, they are supplemented with additional inter-chromosomal contacts. Vertical lines represent multi-way contacts, nodes indicate the corresponding locus' participation in a multi-way contact, and color-coded rows delineate chromosomes. Highlighted boxes indicate example intra-chromosomal contacts (red), inter-chromosomal contacts (magenta), and combinations of intra- and inter-chromosomal contacts (blue). Examples for each type of contact are shown in the top right corner. Multi-way contacts of specific regions are compared between cell types by connecting highlighted boxes with black dashed lines, emphasizing similarities and differences between fibroblasts and B lymphocytes. Normalized degree of loci participating in the top 10 most common multi-way contacts for each chromosome in fibroblast and B-lymphocytes are shown on the left. Red dashed lines indicate the mean degree for fibroblasts and B lymphocytes (top and bottom, respectively). Genomic loci that do not participate in the top 10 most common multi-way contacts for fibroblasts or B lymphocytes were removed from their respective incidence and degree plots. Multi-way contacts were determined in 25 Mb resolution after noise reduction (*Materials and Methods*).

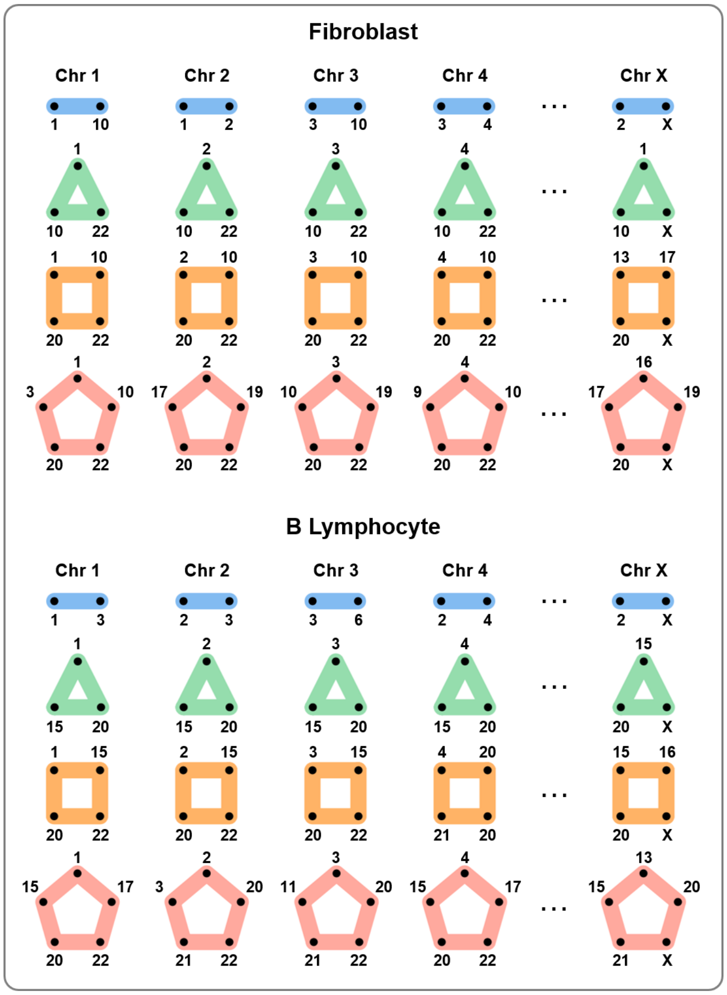

Figure 5: Inter-chromosomal interactions. The most common 2-way, 3-way, 4-way, and 5-way inter-chromosome combinations for each chromosome are represented using motifs from fibroblasts (top, V1-V4) and B lymphocytes (bottom). Rows represent the combinations of 2-way, 3-way, 4-way, and 5-way inter-chromosome interactions, and columns are the chromosomes. Inter-chromosomal combinations are determined using 25 Mb resolution multi-way contacts after noise reduction (*Materials and Methods*) and are normalized by chromosome length. Here we only consider unique chromosome instances (i.e. multiple loci in a single chromosome are ignored).

2.4 Transcription Clusters

Genes are transcribed in short sporadic bursts and transcription occurs in localized areas with high concentrations of transcriptional machinery [10, 11, 12]. This includes transcriptionally engaged polymerase and the accumulation of necessary proteins, called transcription factors. Multiple genomic loci can colocalize at these areas for more efficient transcription. In fact, it has been shown using fluorescence in situ hybridization (FISH) that genes frequently colocalize during transcription [13]. Simulations have also provided evidence that genomic loci which are bound by common transcription factors can self-assemble into clusters, forming structural patterns commonly observed in Hi-C data [12]. We refer to these instances of highly concentrated areas of transcription machinery and genomic loci as *transcription clusters*. The colocalization of multiple genomic loci in transcription clusters naturally leads to multi-way contacts, but these interactions cannot be fully captured from the pairwise contacts of Hi-C. Multi-way contacts derived from Pore-C reads can detect interactions between many genomic loci, and are well suited for identifying potential transcription clusters (Figure 6).

Using the initial criteria of chromatin accessibility and RNA Pol II binding, we identified 16,080 and 16,527 potential transcription clusters from fibroblasts and B lymphocytes, respectively (Table 1, *Materials and Methods*). The majority of these clusters involved at least one expressed gene (72.2% in fibroblasts, 90.5% in B lymphocytes) and many involved at least two expressed genes (31.2% in fibroblasts, 58.7% in B lymphocytes). While investigating the colocalization of expressed genes in transcription clusters, we found that over 30% of clusters containing multiple expressed genes had common transcription factors based on binding motifs (31.0% in fibroblasts, 33.1% in B lymphocytes) and that over half of these common transcription factors were master regulators (56.6% in fibroblasts, 74.7% in B lymphocytes). Two example transcription clusters derived from 3-way, 4-way, and 5-way contacts from both fibroblasts and B lymphocytes are shown in Figure 7. These example clusters contain at least two genes which have at least one common transcription factor.

We tested the criteria for potential transcription clusters for statistical significance (*Materials and Methods)*. That is, we tested whether the identified transcription clusters are more likely to include genes, and if these genes more likely to share common transcription factors, than arbitrary multi-way contacts in both fibroblasts and B lymphocytes. We found that the transcription clusters were significantly more likely to include ≥ 1 gene and ≥ 2 genes than random multi-way contacts ($p < 0.01$). In addition, transcription clusters containing ≥ 2 genes were significantly more likely to have common transcription factors and common master regulators ($p < 0.01$). After testing all order multi-way transcription clusters, we also tested the 3-way, 4-way, 5-way, and 6-way (or more) cases individually. We found that all cases were statistically significant ($p < 0.01$) except for clusters for common transcription factors or master regulators in the 6-way (or more) case for both fibroblasts and B lymphocytes. We hypothesize that these cases were not statistically significant due to the fact that the large number of loci involved in these multi-way contacts will naturally lead to an increase of overlap with genes. This increases the likelihood that at least two genes will have common transcription factors or master regulators. Approximately half of transcription clusters with at least two genes with common transcription factors also contained at least one enhancer locus ($\sim$51% and $\sim$44% in fibroblasts and B lymphocytes, respectively) [14]. This offers even further support that these multi-way contacts represented real transcription clusters.

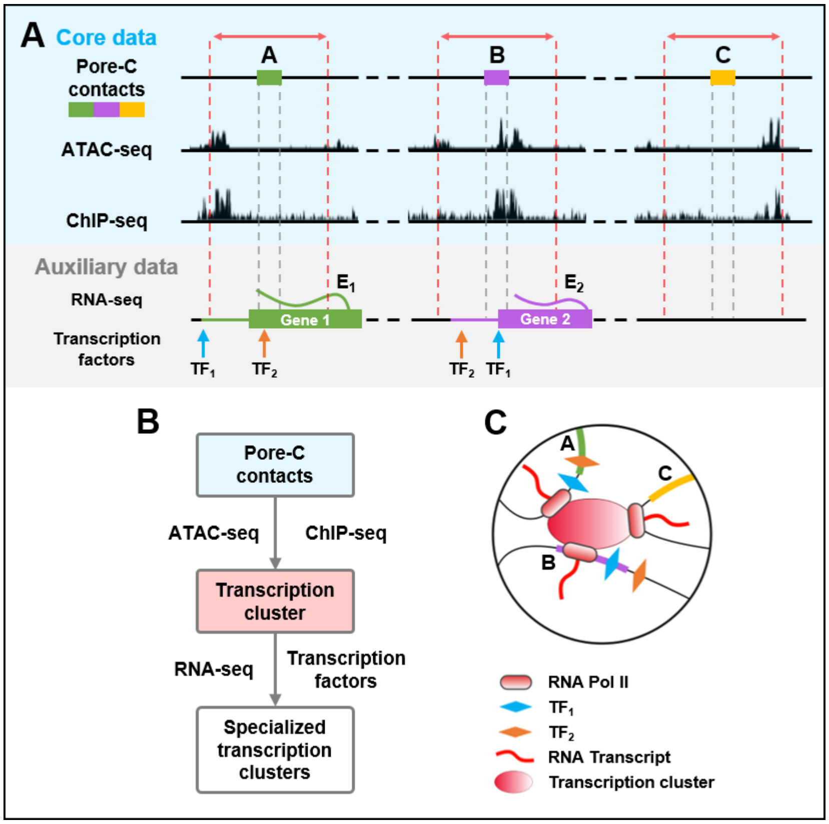

Figure 6: Data-driven identification of transcription clusters. (A) A 5 kb region before and after each locus in a Pore-C read (between red dashed lines) is queried for chromatin accessibility and RNA Pol II binding (ATAC-seq and ChIP-seq, respectively). Multi-way contacts between accessible loci that have ≥ 1 instance of Pol II binding are indicative of potential transcription clusters. Gene expression (RNA-seq, E1 for gene 1 and E2 for gene 2, respectively) and transcription factor binding sites (TF1 and TF2) are integrated to determine potential coexpression and coregulation within multi-way contacts with multiple genes. Transcription factor binding sites are queried ± 5 kb from the gene's transcription start site (Materials and Methods). (B) Pipeline for extracting transcription clusters (Supplementary Materials X). Genes are colored based on the overlapping Pore-C locus, and the extended line from each gene represents the 5 kb flanking region used to query transcription factor binding sites. (C) Schematic representation of a transcription cluster.

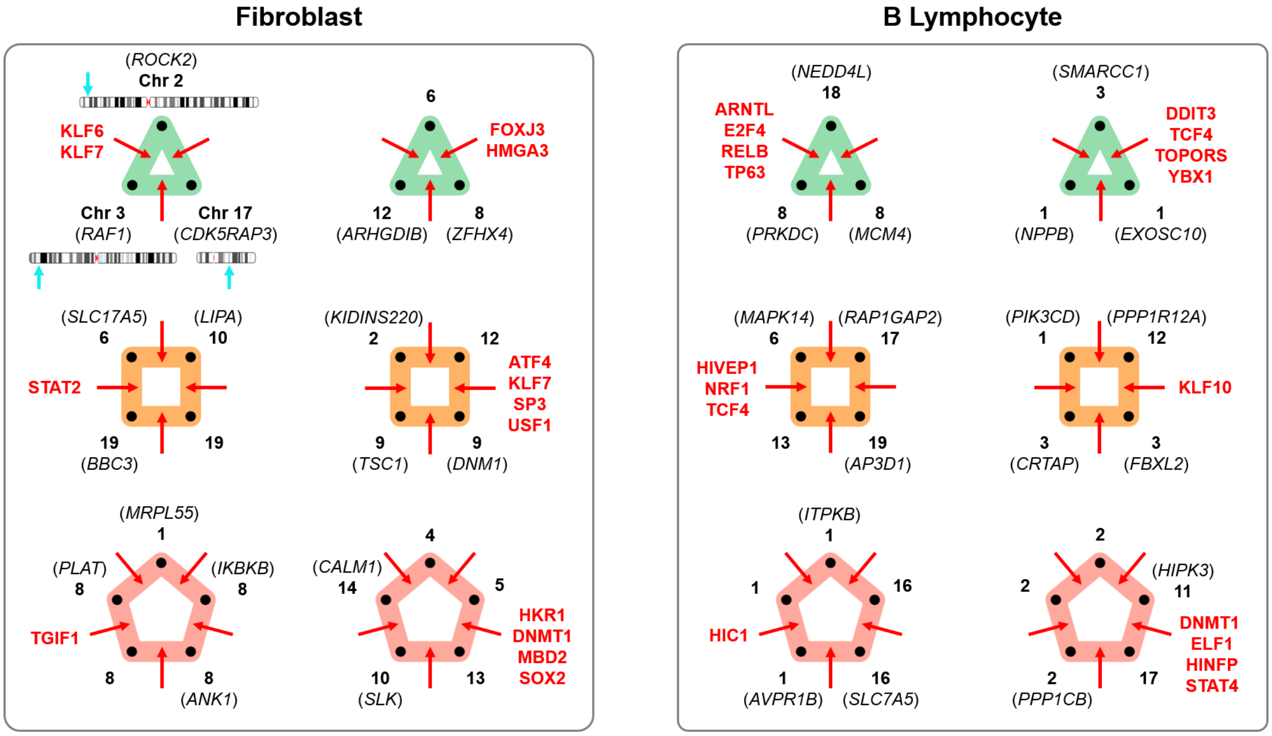

Figure 7: Example transcription clusters. Six examples of potential transcription clusters are shown for fibroblasts (left) and B lymphocytes (right) as multi-way contact motifs. Black labels indicate genes and chromosomes (bold). Red labels correspond to transcription factors shared between ≥ 2 genes within a transcription cluster. Blue arrows indicate a gene's position on its respective chromosome. Multi-way contacts used for fibroblasts include all experiments (V1-V4). Examples were selected from the set of multi-way contacts summarized in the "Clusters with Common TFs" column of Table 1.

Order	Multi-way Contacts	Transcription Clusters	Clusters with $\geq$ 1 Gene	Clusters with $\geq$ 2 Genes	Clusters with Common TFs	Clusters with Common MRs
3	379,165	11,261	7,782	2,986	1,191	679
	240,477	8,384	7,384	4,157	2,006	1,536
4	181,554	3,254	2,519	1,214	276	153
	227,352	4,345	3,972	2,686	822	606
5	98,272	1,021	831	473	63	35
	196,423	1,996	1,881	1,434	277	193
6+	142,575	544	477	341	24	13
	1,000,231	1,802	1,727	1,419	109	67

Table 1: Summary of multi-way contacts. Multi-way contacts from fibroblasts (gray rows, V1-V4) and B lymphocytes (white rows) are listed after different filtering criteria. Multi-way contacts are considered to be potential transcription clusters if all loci within the multi-way contact are accessible and at least one locus has binding from RNA Pol II. These multi-way contacts are then queried for nearby expressed genes. If a transcription cluster candidate has at least two expressed genes, we determine whether the genes have common transcription factors (TFs) through binding motifs. From the set of transcription clusters with common transcription factors, we calculate how many clusters are regulated by at least one master regulator (MR).

Algorithm 1: Multi-way Contact Analysis

```
Input: Aligned Pore-C data ($\mathbf{A}$), RNA-seq ($\mathbf{R}$: gene expression), RNA Pol II ($\mathbf{P}$: ChIP-seq), ATAC-seq ($\mathbf{C}$: chromatin accessibility), transcription factor binding motifs ($\mathbf{B}$)
for each set of Pore-C data $\mathbf{A}_l \in \mathbf{A}$ do
    Construct incidence matrix $\mathbf{H}_l$ using Algorithm 2
    Identify transcription clusters $\mathbf{T}_{lp}$, $\mathbf{T}_{lc}$, and $\mathbf{T}_{ls}$ using Algorithm 3
    Calculate entropy $S_l$ using Algorithm 4
end for
Compute hypergraph distance $d_{ij}$ between pairs $\mathbf{H}_i$ and $\mathbf{H}_j$ with $p \geq 1$ using Algorithm 5
Calculate the statistical significance $\alpha_{ij}$ for hypergraph distance $d_{ij}$ using the permutation test in Algorithm 6.
Return: Hypergraph incidence matrices $\mathbf{H}_l \in \mathbb{R}^{n \times m}$, hypergraph entropy $S_l$, potential transcription clusters $\mathbf{T}_{lp}$, transcription clusters $\mathbf{T}_{lc}$, specialized transcription clusters $\mathbf{T}_{ls}$, and hypergraph distance matrix $[d_{ij}]$ with statistical significance $[\alpha_{ij}]$.
```

3 Discussion

In this work, we introduce a hypergraph framework to study higher-order chromatin structure from long-read sequence data. We demonstrate that multidimensional genomic architecture can be precisely represented and analyzed using hypergraph theory. Hypergraph representations strengthen and extend existing chromatin analysis techniques for study of TADs and intra- and inter-chromosomal interactions. The combination of long-read technology and accurate mathematical representations enable higher fidelity capture of the experimental system and deepen our understanding of genome architecture. Further, using direct capture of multi-way contacts, we identified transcriptional clusters with physical proximity and coordinated gene expression. Our framework thus enables study of explicit structure-function relationships that are observed directly from data, eliminating the need for inference of multi-way contacts. The increased precision of hypergraph representation has the potential to reveal patterns of higher-order differential chromatin organization between multiple cell-types, and further presents the exciting possibility of application at the single cell-level [6, 7, 8, 15].

4 Materials and Methods

Cell culture. Primary human adult dermal fibroblasts were obtained from a donor and were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1X Glutamax (Thermo Fisher Scientific cat no. 35050061) and 1X non-essential amino acid (Thermo Fisher Scientific cat no. 11140050).

Cross-linking. 2.5 million cells were washed three times in chilled 1X phosphate buffered saline (PBS) in a 50 mL centrifuge tube, pelleted by centrifugation at 500 x g for 5 min at 4°C between each wash. Cells were resuspended in 10 mL room temperature 1X PBS 1% formaldehyde (Fisher Scientific cat no. BP531-500) by gently pipetting with a wide bore tip, then incubated at room temperature for 10 min. To quench the cross-linking reaction 527 µL of 2.5 M glycine was added to achieve a final concentration of 1% w/v or 125 mM in 10.5 mL. Cells were incubated for 5 min at room temperature followed by 10 min on ice. The cross-linked cells were pelleted by centrifugation at 500 x g for 5 min at 4°C.

Restriction enzyme digest. The cell pellet was resuspended in 500 µL of cold permeabilization buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% IGEPAL CA-630, 100 µL of protease inhibitor cock-tail Roche cat no. 11836170001) and placed on ice for 15 min. One tablet of protease inhibitor cocktail was dissolved in 1 ml nuclease free water and 100 µL from that was added to a 500 µL permeabilization buffer. Cells were centrifuged at 500 x g for 10 min at 4°C after which the supernatant was aspirated and replaced with 200 µL of chilled 1.5X New England Biolabs (NEB) cutsmart buffer. Cells were centrifuged again at 500 x g for 10 min at 4°C, then aspirated and re-suspended in 300 µL of chilled 1.5X NEB cutsmart buffer. To denature the chromatin, 33.5 µL of 1% w/v sodium dodecyl sulfate (SDS, Invitrogen cat no. 15553-035) was added to the cell suspension and incubated for exactly 10 min at 65°C with gentle agitation then placed on ice immediately afterwards. To quench the SDS, 37.5 µL of 10% v/v Triton X-100 (Sigma Aldrich cat no. T8787-250) was added for a final concentration of 1%, followed by incubation for 10 min on ice. Permeabilized cells were then digested with a final concentration of 1 U/µL of NlaIII (NEB-R0125L) and brought to volume with nuclease-free water to achieve a final 1X digestion reaction buffer in 450 µL. Cells were then mixed by gentle inversion. Cell suspensions were incubated in a thermomixer at 37°C for 18 hours with periodic rotation.

Proximity ligation and reverse cross-linking. NlaIII restriction digestion was heat inactivated at 65°C for 20 min. Proximity ligation was set up at room temperature with the addition of the following reagents: 100 µL of 10X T4 DNA ligase buffer (NEB), 10 µL of 10 mg/mL BSA and 50 µL of T4 Ligase (NEB M0202L) in a total volume of 1000 µL with nuclease-free water. The ligation was cooled to 16°C and incubated for 6 hours with gentle rotation.

Protein degradation and DNA purification. To reverse cross-link, proximity ligated sample was treated with 100 µL Proteinase K (NEB P8107S-800U/ml), 100 µL 10% SDS (Invitrogen cat no. 15553-035) and 500 µL 20% v/v Tween-20 (Sigma Aldrich cat no. P1379) in a total volume of 2000 µL with nuclease-free water. The mixture was incubated in a thermal block at 56°C for 18 hours. In order to purify DNA, the sample was transferred to a 15 mL centrifuge tube, rinsing the original tube with a further 200 µL of nuclease-free water to collect any residual sample, bringing the total sample volume to 2.2 mL. DNA was then purified from the sample using a standard phenol chloroform extraction and ethanol precipitation.

Nanopore sequencing. Purified DNA was Solid Phase Reversible Immobilization (SPRI) size selected before library preparation with a bead ratio of 0.48X for fragments > 1.5 kb. The > 1.5 kb products were prepared for sequencing using the protocol provided by Oxford Nanopore Technologies. In brief, 1 µg of genomic DNA input was used to generate asequencing library according to the protocol provided for the SQK-LSK109 kit. (Oxford Nanopore Technologies, Oxford Science Park, UK). After the DNA repair, end prep, and adapter ligation steps, SPRI select bead suspension (Cat No. B23318, Beckman Coulter Life Sciences, Indianapolis, IN, USA) was used to remove short fragments and free adapters. A bead ratio of 1X was used for DNA repair and end prep while a bead ratio of 0.4X was used for the adapter ligation step. Qubit dsDNA assay (ThermoFisher Scientific, Waltham, MA, USA) was used to quantify DNA and ∼300-400 ng of DNA library was loaded onto a GridION flow cell (version R9, Flo-MIN 106D). In total, 4 sequencing runs were conducted generating a total of 6.25 million reads (referred to as V1-V4).

Sequence processing. Reads which passed Q-score filtering (`--min_qscore 7`, 4.56 million reads) from basecalling on the Oxford Nanopore GridION were used as input for the Pore-C-Snakemake pipeline (`https://github.com/nanoporetech/Pore-C-Snakemake`, commit 6b2f762). The pipeline maps multi-way contacts to a reference genome and stores the hyperedges data in a variety of formats. The reference genome used for mapping was GRCh38.p13 (`https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39/`). The pairs and parquet files output from the pipeline were converted into MATLAB tables to construct hyperedges and the cooler files were used to create the pairwise adjacency matrices (Figures 2 and Supplementary Figure S1). The individual tables from the four sequencing runs were assigned a sequencing run label and then concatenated. The combined tables were used as standard inputs for all downstream software processes.

Hypergraphs. A hypergraph is a generalization of a graph in which its hyperedges can join any number of nodes [16]. Hypergraphs can capture higher-order connectivity patterns and represent multidimensional relationships unambiguously [6, 17]. In our hypergraph representation of genome architecture from Pore-C data, nodes are genomic loci, where a locus can be a gene or a genomic region at a particular resolution (i.e. read level, 100 kb, 1 Mb, or 25 Mb bins). Contacts among loci are represented by hyperedges, where each hyperedge can be one or many contacts. Most higher-order contacts are unique in Pore-C data at high resolution, so for these data we considered unweighted hypergraphs (i.e. ignore the frequency of contacts). For lower resolution (1 Mb or 25 Mb), we considered edge weights (frequency of contacts) to find the most common intra- and inter-chromosomal contacts.

Hypergraph filtering. When binning loci to construct hypergraphs, we performed an additional filtering step based on the frequency of pairwise contacts within multi-way contact data. We first decomposed each multi-way contact into its pairwise combinations at a particular resolution (bin size). From these pairwise contacts, we counted the number of times contact was detected for a pair of loci, and identified the highest frequency locus pairs. Pairwise contacts were kept if detected above a certain threshold number, which was set empirically at the 85^{th} percentile of the most frequently occurring locus pairs. For example, in fibroblast data binned at 1 Mb resolution, a locus pair with six detected contacts corresponded to the 85^{th} percentile. Thus all pairs of loci with fewer than six detected contacts were not considered, which increases confidence in the validity of identified multi-way contacts.

Incidence matrices. An incidence matrix of the genomic hypergraph is an $n \times m$ matrix, where n is the total number of genomic loci, and m is the total number of unique Pore-C contacts (including self-contacts, pairwise contacts, and higher-order contacts). For each column of the incidence matrix, if the genomic locus i is involved in the corresponding Pore-C contact, the ith element of the column is equal to one. If not, it is equal to zero. Thus, nonzero elements in a column show adjacent genomic loci, and the value of nonzero elements is the order of the Pore-C contact. The incidence matrix of the genomic hypergraph can be visualized via PAOHvis [18]. In PAOHvis, genomic loci are parallel horizontal bars, while Pore-C contacts are vertical lines that connects multiple loci (see Figures 1, 2, 3, and 4). Beyond visualization, incidence matrices play a significant role in the mathematical analysis of hypergraphs.

Data-driven identification of transcription clusters. We use Pore-C data in conjunction with multiple other data sources to identify potential transcription clusters (Figure 6). Each locus in a Pore-C read, or multi-way contact, is queried for chromatin accessibility and RNA Pol II binding (ATAC-seq and ChIP-seq peaks, respectively). Multi-way contacts are considered to be potential transcription clusters if all loci within the multi-way contact are accessible and at least one locus has binding from RNA Pol II. These multi-way contacts are then queried for nearby expressed genes. A 5 kb flanking region is added before and after each locus when querying for chromatin accessibility, RNA Pol II binding, and nearby genes [19]. Gene expression (RNA-seq) and transcription factor binding sites are integrated to determine coexpression and coregulation of genes in multi-way contacts. If a transcription cluster candidate has at least two genes present, we determine whether the genes have common transcription factors through binding motifs. From the set of transcription clusters with common transcription factors, we calculate how many clusters are regulated by at least one master regulator, a transcription factor that also regulates its own gene (Figure 6).

Transcription factor binding motifs. Transcription factor binding site motifs were obtained from "The Human Transcription Factors" database [20]. FIMO (https://meme-suite.org/meme/tools/fimo) was used to scan for motifs within ± 5kb of the transcription start sites for protein-coding and microRNA genes. The results were converted to a $22,083 \times 1,007$ MATLAB table, where rows are genes, columns are transcription factors, and entries are the number of binding sites for a particular transcription factor and gene. The table was then filtered to only include entries with three or more binding sites in downstream computations. This threshold was determined empirically and can be adjusted by simple changes to the provided MATLAB code.

Public data sources. Pore-C data for B lymphocytes were downloaded from Ulahannan *et al.* [5]. ATAC-seq and ChIP-seq data were obtained from ENCODE to assess chromatin accessibility and RNA Pol II binding, respectively. These data were compared to read-level Pore-C contacts to determine whether colocalizing loci belong to accessible regions of chromatin and had RNA Pol II binding for both fibroblasts and B lymphocytes. RNA-seq data were also obtained from ENCODE to ensure that genes within potential transcription factories were expressed in their respective cell types. A summary of these data sources can be found in Table 2.

Data Type	Cell Type	Data Description and Source
Pore-C	Fibroblasts	Human primary dermal fibroblasts were derived from a donor skin biopsy
Pore-C	GM12878	B-lymphocyte Pore-C data obtained from Ulahannan *et al.* [5]
ATAC-seq	IMR-90	Fibroblast chromatin accessibility (ENCFF310UDS)
ATAC-seq	GM12878	B-lymphocyte chromatin accessibility data (ENCFF410XEP)
ChIP-seq	IMR-90	Fibroblast RNA Polymerase II binding data (ENCFF676DGR)
ChIP-seq	GM12878	B-lymphocyte RNA Polymerase II binding data (ENCFF912DZY)
RNA-seq	IMR-90	Fibroblast gene expression data averaged over two replicates (ENCFF353SBP, ENCFF496RIW)
RNA-seq	GM12878	B-lymphocyte gene expression data averaged over two replicates (ENCFF306TLL, ENCFF418FIT)
Enhancers	IMR-90	Fibroblast enhancer location data from EnhancerAtlas 2.0 [14]
Enhancers	GM12878	B-lymphocyte enhancer location data from EnhancerAtlas 2.0 [14]

Table 2: Data sources. Data obtained from ENCODE unless otherwise specified [21].

Hypergraph Entropy. Network entropy often is used to measure the connectivity and regularity of a network [22, 23]. We use hypergraph entropy to quantify the organization of chromatin structure from Pore-C data, where higher entropy corresponds to less organized folding patterns (e.g. every genomic locus is highly connected). There are different definitions of hypergraph entropy [8, 24, 25]. In our analysis, we exploit the eigenvalues of the hypergraph Laplacian matrix and fit them into the Shannon entropy formula [25]. In mathematics, eigenvalues can quantitatively represent different features of a matrix [26]. Denote the incidence matrix of the genomic hypergraph by $\mathbf{H}$. The Laplacian matrix then is an n-by-n matrix (n is the total number of genomic loci in the hypergraph), which can be computed by $\mathbf{L} = \mathbf{H}\mathbf{H}^\top \in \mathbb{R}^{n\times n}$, where $\top$ denotes matrix transpose. Therefore, the hypergraph entropy is defined by

$$\textbf{Hypergraph Entropy} = -\sum_{i=1}^{n} \lambda_i \ln \lambda_i, \tag{1}$$

where λ_i are the normalized eigenvalues of $\mathbf{L}$ such that $\sum_{i=1}^{n} \lambda_i = 1$, and the convention $0 \ln 0 = 0$ is used. Biologically, genomic regions with high entropy are likely associated with high proportions of euchromatin, as euchromatin is more structurally permissive than heterochromatin [27, 28].

We computed the entropy of intra-chromosomal genomic hypergraphs for both fibroblasts and B lymphocytes as shown in Figure S2. It is expected that larger chromosomes have larger hypergraph entropy because more potential genomic interactions occur in the large chromosomes. However, there are still subtle differences between the fibroblasts and B lymphocytes chromosomes, indicating differences in their genome structure. In order to better quantify the structure properties of chromosomes and compare between cell types, it may be useful to introduce normalizations to hypergraph entropy in the future.

Hypergraph Distance. Comparing graphs is a ubiquitous task in data analysis and machine learning [29]. There is a rich body of literature for graph distance with examples such as Hamming distance, Jaccard distance, and other spectral-based distances [29, 30, 31]. Here we propose a spectral-based hypergraph distance measure which can be used to quantify global difference between two genomic hypergraphs G_1 and G_2 from two cell lines. Denote the incidence matrices of two genomic hypergraphs by $\mathbf{H}_1 \in \mathbb{R}^{n\times m_1}$ and $\mathbf{H}_2 \in \mathbb{R}^{n\times m_2}$, respectively. For $i = 1, 2$, construct the normalized Lapalacian matrices:

$$\tilde{\mathbf{L}}_i = \mathbf{I} - \mathbf{D}_i^{-\frac{1}{2}}\mathbf{H}_i\mathbf{E}_i^{-1}\mathbf{H}_i^\top\mathbf{D}_i^{-\frac{1}{2}} \in \mathbb{R}^{n\times n}, \tag{2}$$

where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identity matrix, $\mathbf{E}_i \in \mathbb{R}^{m_i\times m_i}$ is a diagonal matrix containing the orders of hyperedges along its diagonal, and $\mathbf{D}_i \in \mathbb{R}^{n\times n}$ is a diagonal matrix containing the degrees of nodes along its diagonal [32]. The degree of a node is equal to the number of hyperedges that contain that node. Therefore, the hypergraph distance between G_1 and G_2 is defined by

$$\textbf{Hypergraph Distance}(\mathsf{G}_1, \mathsf{G}_2) = \frac{1}{n}(\sum_{i=1}^{n} |\lambda_{1j} - \lambda_{2j}|^p)^{\frac{1}{p}}, \tag{3}$$

where λ_{ij} is the jth eigenvalue of $\tilde{\mathbf{L}}_i$ for $i = 1, 2$, and $p \geq 1$. In our analysis, we choose $p = 2$. The hypergraph distance (3) can be used to compare two genomic hypergraphs in a global scale since the eigenvalues of the normalized Laplacian are able to capture global connectivity patterns within the hypergraph.

We computed the distance between the two genome-wide hypergraphs derived from fibroblasts and B lymphocytes, and examined the distance statistically through a permutation test. Figure S3A demonstrates that the two genomic hypergraphs are significantly different, with a p-value < 0.01. Additionally, we computed the distance between intra-chromosomal genomic hypergraphs between fibroblasts and B lymphocytes. We found that Chromosome 19 and 21 have the largest distances between cell types, as seen in Figure S3B.

Statistical Significance via Permutation Test. In order to assess the statistical significance of the transcription cluster candidates we determined using our criteria (Figure 6), we use a permutation test which builds the shape of the null hypothesis (i.e. the random background distribution) by resampling the observed data over N trials. We randomly select n 3rd, 4th, and 5th order multi-way contacts from our Pore-C data, where n is based on the number of transcription cluster candidates we determined for each order using our criteria. For example, we randomly selected $n = 11,261$ multi-way contacts from the set of 3rd order multi-way contacts in fibroblasts (Table 1). For each trial, we determine how many of these randomly sampled "transcription clusters" match our remaining criteria: transcription clusters with ≥ 1 gene, ≥ 2 genes, common TFs, and common MRs. The background distribution for each of the criteria can then be constructed from these values. The proportion of values in these background distributions that are greater than their counterparts from the data-derived transcription cluster candidates yields the p-value. For this analysis, we chose $N = 1,000$ trials. This analysis is based on the assumption that transcription clusters will be more likely to contain genes and that those genes are more likely to have common transcription factors than arbitrary multi-way contacts.

Similarly, we use a permutation test to determine the significance of the measured distances between two hypergraphs. Suppose that we are comparing two hypergraphs G_1 and G_2. We first randomly generate N number of hypergraph $\{\mathsf{R}_i\}_{i=1}^{N}$

that are similar to G_1 ("similar" means similar number of node degree and hyperedge size distribution). The background distribution therefore can be constructed by measuring the hypergraph distances between G_1 and R_i for $i = 1, 2, \ldots, N$. The proportion of distances that are greater than the distance between G_1 and G_2 in this background distribution yields the p-value. For this analysis, we again chose $N = 1,000$ trials.

5 Code and Data Availability

All data generated within this manuscript and MATLAB code in our computational framework can be provided upon request.

6 Acknowledgements

We thank Professor Peter Cook from the University of Oxford for his guidance and helpful discussions on transcription clusters. This work is supported in part by the Air Force Office of Scientific Research (AFOSR) award FA9550-18-1-0028 to IR and the Defense Advanced Research Projects Agency (DARPA) award 140D6319C0020 and the National Science Foundation (NSF) award 2035827 to iReprogram, LLC.

7 Competing Interests

SD is an employee of iReprogram, LLC. LM and IR are members of iReprogram, LLC. NB is an employee of Oxford Nanopore Technologies. SL, CC, and IR have submitted a patent application for the computational framework (2115-008250-US-PS1).

References

[1] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950):289–293, 2009.

[2] Tom Misteli. The self-organizing genome: Principles of genome architecture and function. Cell, 2020.

[3] Haiming Chen, Jie Chen, Lindsey A Muir, Scott Ronquist, Walter Meixner, Mats Ljungman, Thomas Ried, Stephen Smale, and Indika Rajapakse. Functional organization of the human 4d nucleome. Proceedings of the National Academy of Sciences, 112(26):8002–8007, 2015.

[4] Pedro Olivares-Chauvet, Zohar Mukamel, Aviezer Lifshitz, Omer Schwartzman, Noa Oded Elkayam, Yaniv Lubling, Gintaras Deikus, Robert P Sebra, and Amos Tanay. Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature, 540(7632):296–300, 2016.

[5] Netha Ulahannan, Matthew Pendleton, Aditya Deshpande, Stefan Schwenk, Julie M Behr, Xiaoguang Dai, Carly Tyer, Priyesh Rughani, Sarah Kudman, Emily Adney, et al. Nanopore sequencing of dna concatemers reveals higher-order features of chromatin structure. bioRxiv, page 833590, 2019.

[6] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. Science, 353(6295):163–166, 2016.

[7] Can Chen, Amit Surana, Anthony Bloch, and Indika Rajapakse. Controllability of hypergraphs. IEEE Transactions on Network Science and Engineering, 8(2):1646–1657, 2020.

[8] Can Chen and Indika Rajapakse. Tensor entropy for uniform hypergraphs. IEEE Transactions on Network Science and Engineering, 7(4):2889–2900, 2020.

[9] Jie Chen, Alfred O Hero III, and Indika Rajapakse. Spectral identification of topological domains. Bioinformatics, 32(14):2151–2158, 2016.

[10] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. Science, 332(6028):472–474, 2011.

[11] Peter R Cook. The organization of replication and transcription. Science, 284(5421):1790–1795, 1999.

[12] Peter R Cook and Davide Marenduzzo. Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations. Nucleic Acids Research, 46(19):9895–9906, 2018.

[13] Cameron S Osborne, Lyubomira Chakalova, Karen E Brown, David Carter, Alice Horton, Emmanuel Debrand, Beatriz Goyenechea, Jennifer A Mitchell, Susana Lopes, Wolf Reik, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nature Genetics, 36(10):1065–1071, 2004.

[14] Tianshun Gao and Jiang Qian. Enhanceratlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Research, 48(D1):D58–D64, 2020.

[15] Amit Surana, Can Chen, and Indika Rajapakse. Hypergraph dissimilarity measures. arXiv preprint arXiv:2106.08206, 2021.

[16] Claude Berge. Hypergraphs: combinatorics of finite sets, volume 45. Elsevier, 1984.

[17] Michael M Wolf, Alicia M Klinvex, and Daniel M Dunlavy. Advantages to modeling relational data using hypergraphs versus graphs. In 2016 IEEE High Performance Extreme Computing Conference (HPEC), pages 1–7. IEEE, 2016.

[18] Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. IEEE transactions on visualization and computer graphics, 27(1):1–13, 2019.

[19] Shane Neph, Andrew B Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A Stamatoyannopoulos. Circuitry and dynamics of human transcription factor regulatory networks. Cell, 150(6):1274–1286, 2012.

[20] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. Cell, 172(4):650–665, 2018.

[21] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. Nature, 489(7414):57, 2012.

[22] Filippo Passerini and Simone Severini. The von neumann entropy of networks. arXiv:0812.2597, 2008.

[23] Giorgia Minello, Luca Rossi, and Andrea Torsello. On the von neumann entropy of graphs. Journal of Complex Networks, 7(4):491–514, 2019.

[24] Dan Hu, Xue Liang Li, Xiao Gang Liu, and Sheng Gui Zhang. Extremality of graph entropy based on degrees of uniform hypergraphs with few edges. Acta Mathematica Sinica, English Series, 35(7):1238–1250, 2019.

[25] Isabelle Bloch and Alain Bretto. A new entropy for hypergraphs. In International Conference on Discrete Geometry for Computer Imagery, pages 143–154. Springer, 2019.

[26] Gilbert Strang. Introduction to linear algebra, volume 3. Wellesley-Cambridge Press Wellesley, MA, 1993.

[27] Ben D MacArthur and Ihor R Lemischka. Statistical mechanics of pluripotency. Cell, 154(3):484–489, 2013.

[28] Indika Rajapakse, Mark Groudine, and Mehran Mesbahi. What can systems theory of networks offer to biology? PLoS Computational Biology, 8(6):e1002543, 2012.

[29] Claire Donnat and Susan Holmes. Tracking network dynamics: A survey using graph distances. The Annals of Applied Statistics, 12(2):971 – 1012, 2018.

[30] Katherine Faust and John Skvoretz. Comparing networks across space and time, size and species. Sociological Methodology, 32(1):267–299, 2002.

[31] Peter Wills and François G Meyer. Metrics for graph comparison: a practitioner's guide. PloS One, 15(2):e0228728, 2020.

[32] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. Advances in Neural Information Processing Systems, 19:1601–1608, 2006.

8 Supplementary Materials

Algorithm 2: Hypergraph incidence matrix construction

```
Input: Aligned Pore-C data
for each multi-way contact $j$ do
    if multi-way contact contains locus $i$ then
        $\mathbf{H}(i, j) = 1$
    else
        $\mathbf{H}(i, j) = 0$
    end if
end for
Return: Hypergraph incidence matrix $\mathbf{H} \in \mathbb{R}^{n \times m}$ where $n$ is the total number of loci, and $m$ is the total number of
    multi-way contacts.
```

Algorithm 3: Identification of Transcription Clusters

```
Input: Hypergraph incidence matrix $\mathbf{H}$, gene expression $\mathbf{R}$ (RNA-seq), RNA Pol II $\mathbf{P}$ (ChIP-seq), chromatin
    accessibility $\mathbf{C}$ (ATAC-seq), transcription factor binding motifs $\mathbf{B}$
for each multi-way contact $j$ in $\mathbf{H}$ do
    if all loci are accessible from $\mathbf{C}$ and $\geq 1$ locus has Pol II binding from $\mathbf{P}$ then
        multi-way contact $j$ from $\mathbf{H}$ is added to the set of potential transcription clusters $\mathbf{T}_p$
    end if
end for
for each potential transcription cluster $k$ in $\mathbf{T}_p$ do
    if loci contain $\geq 2$ expressed genes from $\mathbf{R}$ which have $\geq 1$ common TFs from $\mathbf{B}$ then
        multi-way contact $k$ from $\mathbf{T}_p$ is added to the set of transcription clusters $\mathbf{T}_c$
    end if
    if loci contain $\geq 2$ expressed genes from $\mathbf{R}$ which have $\geq 1$ common MRs from $\mathbf{B}$ then
        multi-way contact $k$ from $\mathbf{T}_p$ is added to the set of transcription clusters $\mathbf{T}_s$
    end if
end for
Return: Potential transcription clusters $\mathbf{T}_p$, transcription clusters $\mathbf{T}_c$, and specialized transcription clusters $\mathbf{T}_s$
```

Algorithm 4: Hypergraph Entropy [25]

```
Input: Hypergraph incidence matrix $\mathbf{H} \in \mathbb{R}^{n \times m}$
Construct the hypergraph Laplacian matrix $\mathbf{L} = \mathbf{H}\mathbf{H}^\top \in \mathbb{R}^{n \times n}$
Compute the eigenvalues $\lambda_i$ of $\mathbf{L}$ using eigendecomposition
Normalize the eigenvalues $\bar{\lambda}_j = \frac{\lambda_j}{\sum_{i=1}^{n} \lambda_i}$
Compute the hypergraph entropy
    $S = -\sum_{j} \bar{\lambda}_j \ln \bar{\lambda}_j$
Return: Hypergraph entropy $S$.
```

Algorithm 5: Comparing Hypergraphs

```
Input: Two hypergraph $\mathsf{G}_1$ and $\mathsf{G}_2$ with incidence matrices $\mathbf{H}_1 \in \mathbb{R}^{n\times m_1}$ and $\mathbf{H}_2 \in \mathbb{R}^{n\times m_2}$
Construct the normalized hypergraph Laplacian matrices
    $\tilde{\mathbf{L}}_i = \mathbf{I} - \mathbf{D}_i^{-\frac{1}{2}}\mathbf{H}_i\mathbf{E}_i^{-1}\mathbf{H}_i^\top\mathbf{D}_i^{-\frac{1}{2}} \in \mathbb{R}^{n\times n}$,
  where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identity matrix, $\mathbf{E}_i \in \mathbb{R}^{m_i\times m_i}$ is a diagonal matrix containing the orders of hyperedges along its diagonal, and $\mathbf{D}_i \in \mathbb{R}^{n\times n}$ is a diagonal matrix containing the degrees of nodes along its diagonal, for $i = 1, 2$ [32].
Compute the hypergraph distance
    $d = \frac{1}{n}\left(\sum_{i=1}^{n} |\lambda_{1j} - \lambda_{2j}|^p\right)^{\frac{1}{p}}$,
  where $\lambda_{ij}$ is the $j$th eigenvalue of $\tilde{\mathbf{L}}_i$ for $i = 1, 2$, and $p \geq 1$.
Return: Hypergraph distance $d$ between $\mathsf{G}_1$ and $\mathsf{G}_2$.
```

Algorithm 6: Permutation Test for Hypergraph Distance

```
Input: Two hypergraph $\mathsf{G}_1$ and $\mathsf{G}_2$, a prescribed significant level $\alpha$
Null hypothesis $H_0$: $\mathsf{G}_1$ and $\mathsf{G}_2$ are similar, and alternative hypothesis $H_1$: $\mathsf{G}_1$ and $\mathsf{G}_2$ are dissimilar
Randomly generate $N$ number of hypergraph $\{\mathsf{R}_i\}_{i=1}^N$ that are similar to $\mathsf{G}_1$ ("similar" means similar number of node degree and hyperedge size distribution)
Construct the the background distribution by measuring the hypergraph distances between $\mathsf{G}_1$ and $\mathsf{R}_i$ for $i = 1, 2, \dots, N$
Compute the actual hypergraph distance between $\mathsf{G}_1$ and $\mathsf{G}_2$
Obtain the p-value by calculating the proportion of distances that are greater than the actual distance in the background distribution
if p-value $\leq \alpha$ then
  Reject the null hypothesis $H_0$
end if
Return: Permutation test result of whether $\mathsf{G}_1$ and $\mathsf{G}_2$ are similar or dissimilar.
```

Algorithm 7: Identification of Transcription Clusters with Enhancers

```
Input: Hypergraph incidence matrix $\mathbf{H}$, gene expression $\mathbf{R}$ (RNA-seq), RNA Pol II $\mathbf{P}$ (ChIP-seq), chromatin accessibility $\mathbf{C}$ (ATAC-seq), enhancer locations $\mathbf{E}$, transcription factor binding motifs $\mathbf{B}$
for each multi-way contact $j$ in $\mathbf{H}$ do
  if all loci are accessible from $\mathbf{C}$ and $\geq 1$ locus has Pol II binding from $\mathbf{P}$ then
    multi-way contact $j$ from $\mathbf{H}$ is added to the set of potential transcription clusters $\mathbf{T}_p$
  end if
end for
for each potential transcription cluster $k$ in $\mathbf{T}_p$ do
  if loci contain $\geq 2$ expressed genes from $\mathbf{R}$ which have $\geq 1$ common TFs from $\mathbf{B}$ and $\geq 1$ enhancer from $\mathbf{E}$ then
    multi-way contact $k$ from $\mathbf{T}_p$ is added to the set of transcription clusters $\mathbf{T}_c$
  end if
  if loci contain $\geq 2$ expressed genes from $\mathbf{R}$ which have $\geq 1$ common MRs from $\mathbf{B}$ and $\geq 1$ enhancer from $\mathbf{E}$ then
    multi-way contact $k$ from $\mathbf{T}_p$ is added to the set of transcription clusters $\mathbf{T}_s$
  end if
end for
Return: Potential transcription clusters $\mathbf{T}_p$, transcription clusters $\mathbf{T}_c$, and specialized transcription clusters $\mathbf{T}_s$
```

9 Supplementary Figures

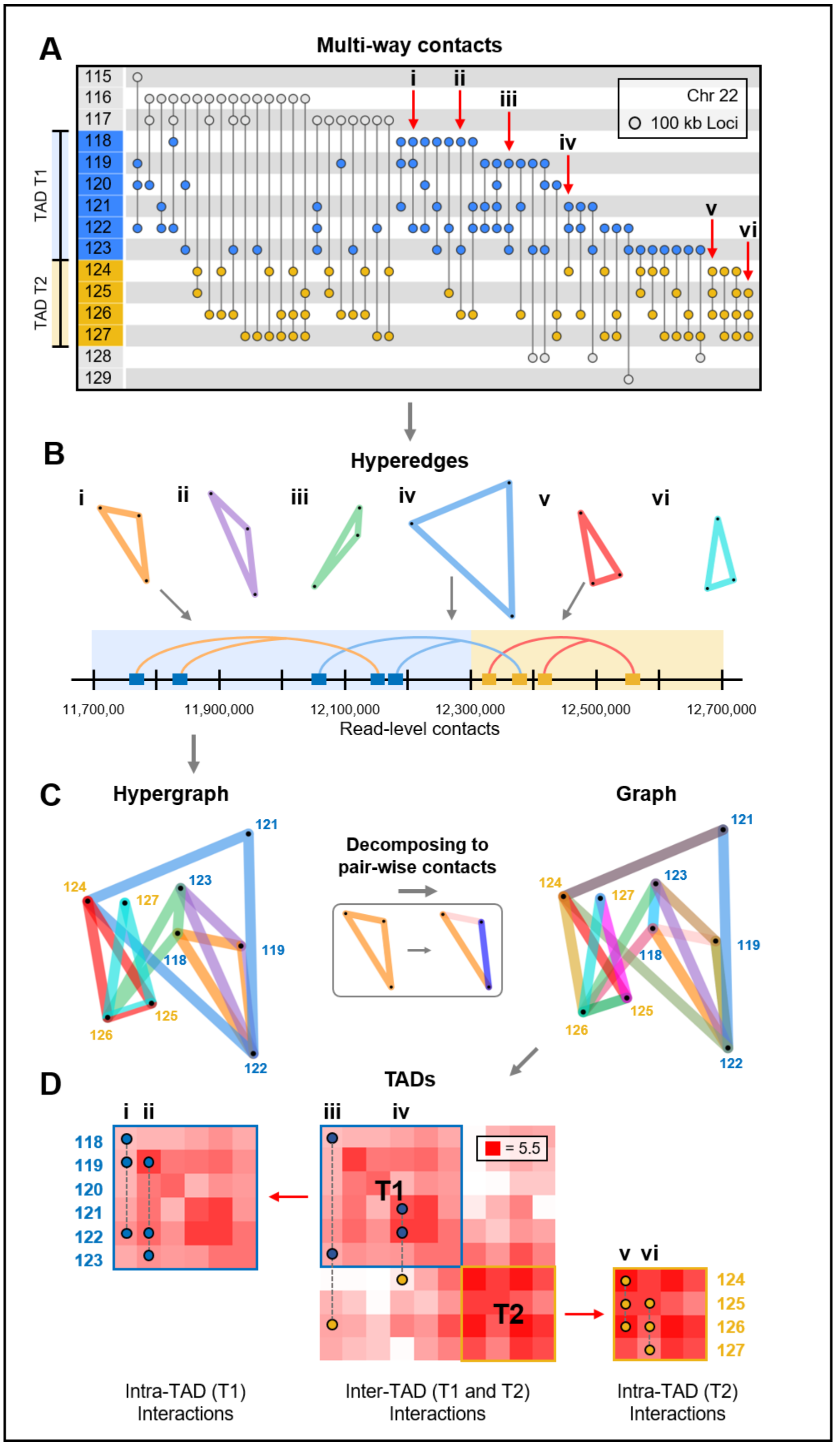

Figure S1: Local organization of the genome. (A) Incidence matrix visualization of a region in Chromosome 22 from B lymphocytes. The numbers in the left column represent genomic loci, vertical lines represent multi-way contacts, where nodes indicate the corresponding locus' participation in this contact. The blue and yellow regions represent two TADs, T1 and T2. The six contacts, denoted by the labels i-vi, are used as examples for hypergraph and genomic folding pattern visualizations. (B) Hypergraph visualization of the multi-way contacts i-vi from A. Blue and yellow labels indicate which TADs these loci participate in. (C) Contact frequency matrices were constructed by separating all multi-way contacts within and between the two TADs into their pairwise combinations. Example multi-way contacts are superimposed onto contact frequency matrices. All multi-way contacts in this figure were determined in 100 kb resolution after noise reduction (*Materials and Methods*).

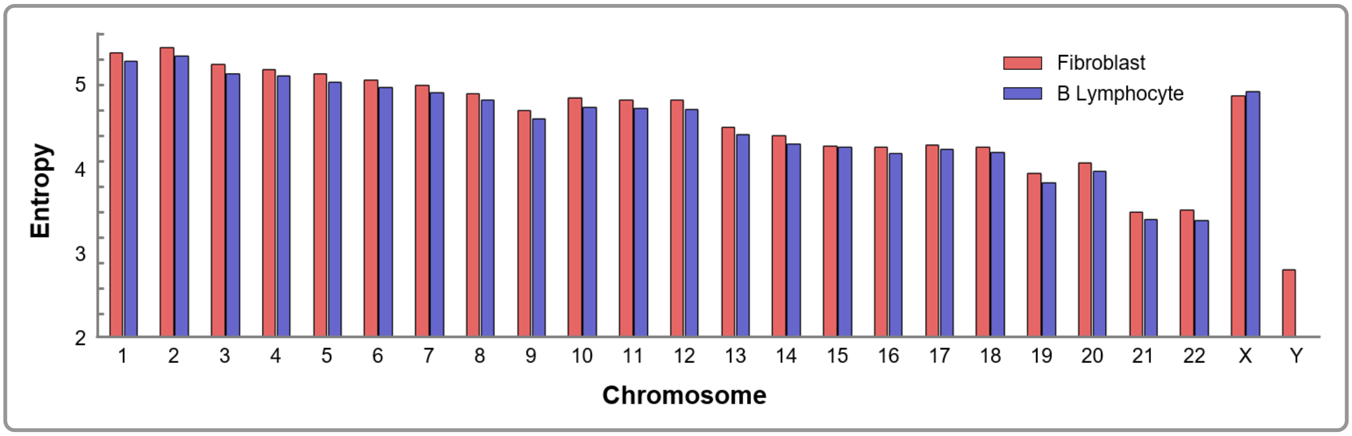

Figure S2: Entropy of intra-chromosomal genomic hypergraph for fibroblast and B lymphocytes.

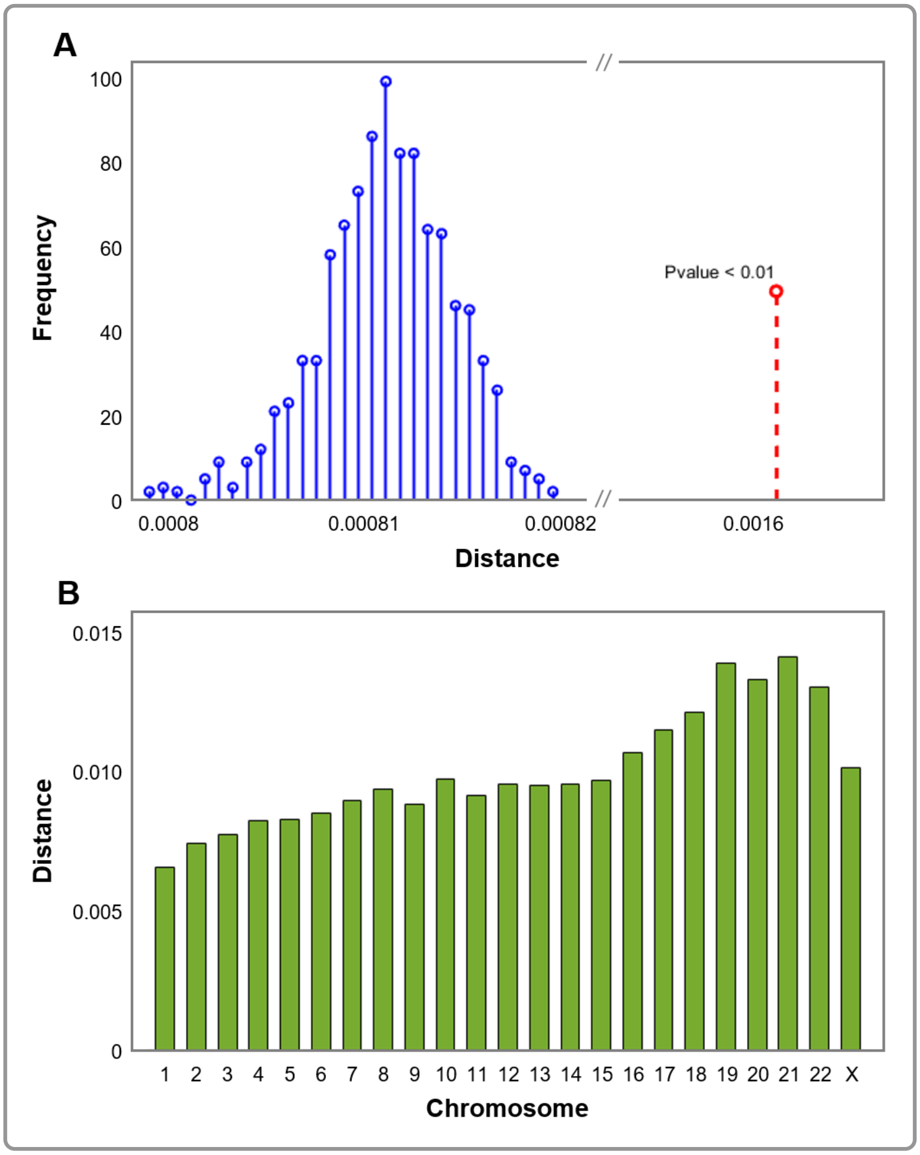

Figure S3: Hypergraph distance between two genome-wide hypergraphs derived from fibroblast and B lymphocytes. (A) Background distribution formed by measuring the hypergraph distances between the hypergraph derived from fibroblast and random hypergraphs. The actual distance between two genome-wide hypergraphs is also highlighted in red with a p-value ¡ 0.01. (B) Hypergraph distances between intra-chromosomal genomic hypergraphs between fibroblasts and B lymphocytes.

10 Supplementary Tables

File Type	Description
.fastq	Text file which contains unique identifiers for Pore-C reads and the raw sequences contained within each read
.pairs	Text file which contains pairs of aligned genomic loci at base-pair resolution, which can be grouped together using unique identifiers to construct multi-way contacts
.parquet	Text file which contains aligned multi-way contacts at base-pair resolution, restriction fragment assignments, and indicators for the quality of read alignment
.mcool	Binary storage file which contains pair-wise interactions from Pore-C data at multiple resolutions

Table S1: Descriptions of file types used within the computational framework (Figure 1).

Order	Multi-way Contacts	Transcription Clusters	Clusters with $\geq$ 2 Genes	Clusters with $\geq$ 1 Enhancer	Clusters with Common TFs	Clusters with Common MRs
3	379,165	11,261	2,986	2,914	715	336
	240,477	8,384	4,157	3,487	1,092	739
4	181,554	3,254	1,214	1,208	80	34
	227,352	4,345	2,686	2,387	270	168
5	98,272	1,021	473	467	8	6
	196,423	1,996	1,434	1,315	53	32
6+	142,575	544	341	341	1	0
	1,000,231	1,802	1,419	1,343	7	2

Table S2: Summary of multi-way contacts with enhancers. Multi-way contacts from fibroblasts (gray rows, V1-V4) and B lymphocytes (white rows) are listed after different filtering criteria. Multi-way contacts are considered to be potential transcription clusters if all loci within the multi-way contact are accessible and at least one locus has binding from RNA Pol II. These multi-way contacts are then queried for nearby expressed genes. If a transcription cluster candidate has at least two expressed genes and at least one enhancer locus, we determine whether the genes have common transcription factors (TFs) through binding motifs. From the set of transcription clusters with common transcription factors, we calculate how many clusters are regulated by at least one master regulator (MR).

Chr	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10+**
1	230	8,333	14,492	5,501	2,093	872	353	133	54	57
2	241	9,183	15,548	6,180	2,373	873	437	197	71	95
3	198	7,113	13,631	5,113	1,909	750	302	132	57	50
4	190	6,342	11,147	4,360	1,642	598	284	113	56	28
5	179	5,785	10,059	3,897	1,457	640	249	106	39	31
6	167	5,361	9,368	3,551	1,364	511	207	75	39	27
7	159	4,851	8,365	3,258	1,274	453	180	72	35	26
8	143	4,321	7,596	2,864	1,069	393	154	71	32	16
9	122	2,722	4,899	1,831	614	209	76	31	8	6
10	134	4,010	7,185	2,695	912	337	144	63	23	14
11	133	3,698	6,613	2,533	890	330	101	62	14	14
12	132	3,815	6,630	2,497	878	324	114	47	17	12
13	97	2,627	4,767	1,751	588	247	82	37	11	8
14	87	2,209	3,932	1,366	509	140	51	19	10	2
15	82	1,623	2,812	991	298	98	29	16	2	1
16	80	1,650	3,126	1,181	363	127	39	11	5	2
17	82	1,345	2,374	732	201	66	20	4	3	2
18	77	1,828	3,230	1,128	368	149	51	16	10	4
19	56	873	1,455	462	118	41	7	3	0	0
20	64	1,247	2,238	783	222	69	27	6	3	0
21	39	453	809	264	79	18	6	1	0	0
22	39	438	728	206	43	13	3	0	1	0
X	151	1,078	1,852	882	350	119	43	17	10	0
Y	23	70	93	25	6	4	0	0	0	0

Table S3: Fibroblast intra-chromosome contact orders (1 Mb resolution).

Chr	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10+**
1	232	4,725	17,254	16,396	10,752	6,836	4,248	2,663	1,652	3,130
2	243	4,964	16,695	16,702	11,787	7,891	5,172	3,432	2,288	4,631
3	199	4,055	16,351	14,795	9,846	6,376	4,023	2,691	1,759	3,293
4	191	3,725	13,399	13,309	9,192	6,218	4,081	2,632	1,698	3,301
5	180	3,449	12,352	12,458	8,799	5,810	3,672	2,358	1,548	2,728
6	171	3,300	11,249	11,640	7,862	5,307	3,235	2,139	1,376	2,486
7	160	3,086	11,350	11,272	7,758	5,101	3,223	2,213	1,391	2,613
8	145	2,895	10,755	10,608	7,333	4,778	3,141	2,068	1,289	2,477
9	124	2,399	9,360	8,535	5,847	3,802	2,393	1,572	931	1,753
10	134	2,580	9,429	9,273	6,203	4,113	2,576	1,632	1,020	1,856
11	134	2,505	8,979	8,661	5,795	3,629	2,250	1,408	852	1,530
12	134	2,376	8,329	8,200	5,465	3,449	2,192	1,344	833	1,304
13	99	1,763	6,665	6,527	4,585	2,935	1,834	1,078	635	1,159
14	89	1,531	5,909	5,884	3,863	2,236	1,430	858	514	843
15	85	1,349	5,064	5,130	3,354	1,964	1,164	706	372	582
16	83	1,444	5,993	5,722	3,769	2,437	1,419	808	494	813
17	84	1,355	4,278	4,294	2,612	1,514	882	475	271	366
18	80	1,478	6,015	5,569	3,616	2,359	1,380	861	549	840
19	57	884	3,335	3,251	1,998	1,141	664	383	224	252
20	65	1,053	4,344	4,149	2,684	1,656	902	547	315	441
21	40	481	2,285	2,245	1,392	747	402	235	96	127
22	39	445	1,739	1,795	1,098	512	255	127	59	72
X	156	3,681	12,552	11,686	8,206	5,915	3,941	2,738	1,950	4,414

Table S4: GM12878 intra-chromosome contact orders (1 Mb resolution).